Friday, January 22. 2010

Less known Solaris features - IP Multipathing (Part 7): New IPMP and link aggregation

As I wrote before there is another way to protect your system against the failure of a network connection - Link Aggregation. As i've explained before, there are failure modes that can't be addressed by link aggregation. But you can use both in conjunction. This makes sense, when your main connection is a 10GBe interface and you don't want to plug a second one into the system and use already existent 1GBe Interfaces as a backup for it instead.
It's pretty straightforward to do so. At first you have to configure the link aggregation.

```
jmoekamp@hivemind:~$ pfexec bash
jmoekamp@hivemind:~# ifconfig e1000g0 unplumb
jmoekamp@hivemind:~# ifconfig e1000g1 unplumb
jmoekamp@hivemind:~# dladm create-aggr -l e1000g0 -l e1000g1 aggregate0
jmoekamp@hivemind:~# dladm show-aggr -x aggregate0
LINK        PORT         SPEED DUPLEX   STATE    ADDRESS           PORTSTATE
aggregate0  --           0Mb  unknown   unknown  0:1b:21:3d:91:f7  --
            e1000g0      0Mb  half      down     0:1b:21:3d:91:f7  standby
            e1000g1      0Mb  half      down     0:1b:21:16:8d:7f  standby
```

The dladm create-aggr creates an aggregation, that bundles the interfaces e1000g0 and e1000g1 into a single virtual interface. Now I plug both cables into the switch.

```
jmoekamp@hivemind:~# dladm show-aggr -x aggregate0
LINK        PORT         SPEED DUPLEX   STATE    ADDRESS           PORTSTATE
aggregate0  --           100Mb full     up       0:1b:21:3d:91:f7  --
            e1000g0      100Mb full     up       0:1b:21:3d:91:f7  attached
            e1000g1      100Mb full     up       0:1b:21:16:8d:7f  attached
```

Interfaces are up, the aggregation is ready for use.

```
jmoekamp@hivemind:~# ifconfig rge0 unplumb
jmoekamp@hivemind:~# ifconfig production0 ipmp hivemind-prod up
jmoekamp@hivemind:~# ifconfig aggregate0 plumb
jmoekamp@hivemind:~# ifconfig aggregate0 -failover group production0 up
jmoekamp@hivemind:~# ifconfig rge0 plumb
jmoekamp@hivemind:~# ifconfig rge0 -failover group production0 upLooks pretty much like a standard IPMP
```

configuration. You can think of aggregate0 as a plain-standard physical interface from the perspective the the admin. When we check the IPMP configuration we will see both interfaces.

```
jmoekamp@hivemind:~# ipmpstat -i
INTERFACE   ACTIVE  GROUP      FLAGS     LINK      PROBE     STATE
rge0        yes     production0 -------   up        disabled  ok
aggregate0  yes     production0 --mb---   up        disabled  ok
jmoekamp@hivemind:~# ipmpstat -g
GROUP       GROUPNAME   STATE     FDT       INTERFACES
production0 production0 ok        --        rge0 aggregate0Now we unplug one of the aggregated cables.
jmoekamp@hivemind:~# dladm show-aggr -x aggregate0
LINK        PORT         SPEED DUPLEX   STATE    ADDRESS           PORTSTATE
aggregate0  --           100Mb full     up       0:1b:21:3d:91:f7  --
            e1000g0      100Mb full     up       0:1b:21:3d:91:f7  attached
            e1000g1      0Mb  half      down     0:1b:21:16:8d:7f  standby
jmoekamp@hivemind:~# ipmpstat -g
GROUP       GROUPNAME   STATE     FDT       INTERFACES
production0 production0 ok        --        rge0 aggregate0
jmoekamp@hivemind:~# ipmpstat -i
INTERFACE   ACTIVE  GROUP      FLAGS     LINK      PROBE     STATE
rge0        yes     production0 -------   up        disabled  ok
aggregate0  yes     production0 --mb---   up        disabled  ok
```

Everything is still okay. The aggregate hides the fact of the one failed interface from the IPMP subsystem. Now we unplug the second interface.

```
jmoekamp@hivemind:~# ipmpstat -i
INTERFACE   ACTIVE  GROUP      FLAGS     LINK      PROBE     STATE
rge0        yes     production0 --mb---   up        disabled  ok
aggregate0  no      production0 -------   down      disabled  failed
```

```
jmoekamp@hivemind:~# ipmpstat -g
GROUP      GROUPNAME  STATE    FDT      INTERFACES
production0 production0 degraded  --       rge0 [aggregate0]
jmoekamp@hivemind:~# ipmpstat -i
INTERFACE  ACTIVE  GROUP      FLAGS    LINK     PROBE    STATE
rge0       yes     production0 --mb---  up       disabled  ok
aggregate0 no      production0 -------  down     disabled  failed
jmoekamp@hivemind:~# dladm show-aggr -x aggregate0
LINK        PORT          SPEED DUPLEX   STATE    ADDRESS           PORTSTATE
aggregate0  --            0Mb  unknown   down     0:1b:21:3d:91:f7  --
            e1000g0       0Mb  half      down     0:1b:21:3d:91:f7  standby
            e1000g1       0Mb  half      down     0:1b:21:16:8d:7f  standby
```

The links are both down, and without a functional interface left, the "link" of the aggregate goes down as well (It stays up, as long as there's a functional interface in the aggregate). Of course the IPMP subsystem switches to rge0 now. When we plug one cable back to the switch, the aggregate is functional again and IPMP detects this and the interface is considered as functional in IPMP again, too.

```
jmoekamp@hivemind:~# dladm show-aggr -x aggregate0
LINK        PORT          SPEED DUPLEX   STATE    ADDRESS           PORTSTATE
aggregate0  --            100Mb full     up       0:1b:21:3d:91:f7  --
            e1000g0       100Mb full     up       0:1b:21:3d:91:f7  attached
            e1000g1       0Mb  half      down     0:1b:21:16:8d:7f  standby
jmoekamp@hivemind:~# ipmpstat -i
INTERFACE  ACTIVE  GROUP      FLAGS    LINK     PROBE    STATE
rge0       yes     production0 --mb---  up       disabled  ok
aggregate0 yes     production0 -------  up       disabled  ok
```

When you plug the second interface into the interface, the aggregate is complete. But it doesn't change a thing from the IPMP side, as the aggregate0 interface was already functional from the perpective of IPMP with just one interface.

```
jmoekamp@hivemind:~# dladm show-aggr -x aggregate0
LINK        PORT          SPEED DUPLEX   STATE    ADDRESS           PORTSTATE
aggregate0  --            100Mb full     up       0:1b:21:3d:91:f7  --
            e1000g0       100Mb full     up       0:1b:21:3d:91:f7  attached
            e1000g1       100Mb full     up       0:1b:21:16:8d:7f  attached
jmoekamp@hivemind:~#
```

Posted by Joerg Moellenkamp in English, Oracle, Solaris at 10:59